Am. J. Hum. Genet. 73:801–811, 2003

Using the Noninformative Families in Family-Based Association Tests: A Powerful New Testing Strategy

Christoph Lange,[1] Dawn DeMeo,[2] Edwin K. Silverman,[2] Scott T. Weiss,[2] and Nan M. Laird[1]

[1]Department of Biostatistics, Harvard School of Public Health, and [2]Channing Laboratory, Harvard Medical School, Boston

For genetic association studies with multiple phenotypes, we propose a new strategy for multiple testing with family-based association tests (FBATs). The strategy increases the power by both using all available family data and reducing the number of hypotheses tested while being robust against population admixture and stratification. By use of conditional power calculations, the approach screens all possible null hypotheses without biasing the nominal significance level, and it identifies the subset of phenotypes that has optimal power when tested for association by either univariate or multivariate FBATs. An application of our strategy to an asthma study shows the practical relevance of the proposed methodology. In simulation studies, we compare our testing strategy with standard methodology for family studies. Furthermore, the proposed principle of using all data without biasing the nominal significance in an analysis prior to the computation of the test statistic has broad and powerful applications in many areas of family-based association studies.

Introduction

In many studies of genetic association between phenotypes and genetic markers, samples of subjects, along with their parents or other family members, are recorded. Family-based association tests (FBATs) (Spielman et al. 1993; Thomson 1995; Zhao 2000; Laird 2000) can then be constructed by using the genetic data of family members to derive the distribution of a test statistic under the null hypothesis, conditioning on the observed phenotypes (Rabinowitz and Laird 2000).

FBATs can be powerful tests for linkage between a marker and a disease-susceptibility locus, in the presence of linkage disequilibrium between the two loci (Risch and Merikangas 1996). In studies of complex disorders, one may record a large number of phenotypes related to the disorder. However, the phenotype with the strongest genetic component attributable to the tested marker is often not known prior to the analysis, making it desirable to test all recorded phenotypes. This can lead to many tests and therefore requires a correction for multiple testing. Standard adjustments for multiple testing (e.g., Bonferroni correction or Hochberg correction) become severe when the number of tests is large (i.e., the significance levels for the individuals tests become unrealistically small). Although these methods are easily applied, they are unrealistically conservative for many applications. In this setting, one may fail to establish an overall significance between the phenotypes and the marker locus of interest.

DeMeo et al. (2002) and Lange et al. (2003) suggested reducing the number of null hypotheses by testing several phenotypes simultaneously. They grouped phenotypes into symptom groups and tested all phenotypes of one symptom group simultaneously by a single multivariate test. They found significant associations that could not be detected by univariate tests. However, it is not always obvious how one should define such symptom groups. DeMeo et al. relied on their "clinical intuition" for their definition of the phenotype/symptom groups, but this approach can be difficult when there is little knowledge about the joint genetic components of the phenotypes, and the results may not be reproducible by other investigators.

In the present article, we propose a systematic approach to construct the most powerful FBAT statistic for scenarios in which the genetic data are given and multiple phenotypes have been recorded. The distribution of FBATs under the null hypothesis is computed by conditioning on the parental information and on Mendel's law of random segregation. Because the transmission from a homozygous parent is not random, offspring with homozygous parents do not contribute to the FBAT statistic. In the present article, we will call families with two homozygous parents "noninformative families," and families with at least one heterozygous parent will be referred to as "informative." Our algorithm for the construction of the most powerful FBAT statistic can be divided into six steps. We repeat the first

Received April 16, 2003; accepted for publication July 14, 2003; electronically published September 18, 2003.

Address for correspondence and reprints: Dr. Christoph Lange, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115. E-mail: clange@hsph.harvard.edu

0002-9297/2003/7304-0009$15.00

five steps of the algorithm for all possible subsets of phenotypes, and, in the sixth step, we select the subset of phenotypes for which the power is maximal. The steps of the algorithm are as follows: (1) Select any subset of phenotypes. (2) Posit a multivariate model that describes the selected phenotypes as a function of the genotypes. (3) Because the use of the observed offspring genotypes in the informative families would bias the nominal significance level of the FBAT statistic, replace the observed offspring genotypes in the multivariate model by their expected values conditional on the parental genotypes or the sufficient statistic (Rabinowitz and Laird 2000). (4) For this adjusted multivariate model, estimate the effect-size parameters, using the generalized estimating equation approach of Liang and Zeger (1986), which is robust against misspecification of the environmental variance. (5) Using the approach to conditional power calculation described by Lange and Laird (2002*b*), estimate the power for the selected phenotypes when tested simultaneously by the FBAT–generalized estimating equation (GEE) (Lange et al. 2003). (6) Use the multivariate FBAT-GEE on the subset of phenotypes with maximal power.

This strategy is entirely robust against population admixture and stratification, since the decision regarding a potential association is based solely on the FBAT-GEE statistic, which is robust against these effects. However, population admixture and stratification that are not accounted for in the multivariate model for the phenotypes will affect the performance of the phenotype selection. Consequently, population admixture and stratification will have an influence on the power of our testing strategy, but the strategy itself remains robust against these effects. Since our approach utilizes the data on all available families to estimate the effect size of the tests, it is far more powerful than standard univariate or multivariate FBATs, even in the presence of population admixture and stratification. The practical relevance and the robustness of our approach will be further illustrated by application to an asthma study (DeMeo et al. 2002). Using simulation studies, we compare our approach with other approaches to multiple testing, in the presence and absence of population admixture and stratification and for ascertained samples.

It is important to note that the approach discussed here can be applied only when there is variation in the phenotypes. For data sets obtained through a strong ascertainment condition for the analyzed phenotypes (i.e., only affected offspring are ascertained), the effect-size estimates must be obtained from an independent sample with phenotypic variation. The approach will be outlined for quantitative traits. When other types of phenotypes are observed (e.g., counts or dichotomous variables), generalized linear models (McCullagh and Nelder 1989) can be used for the effect-size estimates.

Methods

In this section, we explain the methodology used to estimate the power of FBATs for a given data set without biasing the nominal significance level. In principle, the proposed methodology must be applied to all possible subsets of phenotypes, to obtain the subset with the best result. To keep the derivations and equations simple, we assume that a biallelic marker with alleles A and B is given and that the marker locus is the disease locus. The allele frequency of the disease gene will be denoted by p. Furthermore, n independent families are sampled, and the ith family has m_i offspring. The number of transmitted A alleles for the jth offspring in the ith family is given by X_{ij}, with $X_{ij} = 0,1,2$. For each offspring, K traits are recorded and denoted by $y_{ij1},\ldots,y_{ijk}$. The K-dimensional vector of phenotypes for the jth offspring in the ith family is defined by $y_{ij}^t = (y_{ij1},\ldots,y_{ijk})$. In the present article, we assume that the parental genotypes are observed and are denoted in the ith family by p_{i_1} and p_{i_2}. However, the proposed methodology extends readily to scenarios in which parental information is missing, as outlined by Laird et al. (2000).

For simplicity, we assume that the effect of the underlying quantitative trait loci (QTL) is additive. Then the standard quantitative genetic model (Falconer and Mackay 1997) for the phenotypic mean is given by

$$E(Y_{ijk}) = \mu_k + a_k x_{ij} , \qquad (1)$$

where μ_k is the overall mean for the kth phenotype and a_k is the additive effect size for the kth phenotype. Denoting the vector of phenotypes for the siblings in the ith family by $\mathbf{y}_i = (y_{i_1}^t,\ldots,y_{im_1}^t)$, the phenotypic variance for the ith family is defined by

$$\mathrm{Var}(\mathbf{Y}_i) = \mathbf{V}_i , \qquad (2)$$

where $\mathbf{V}_i$ is a $(Km_i \times Km_i)$ variance matrix with components that are attributable to the putative QTL and to shared environmental and polygenic effects (Abecasis et al. 2000).

Most power calculations are hypothetical, in that the effect a_k is prespecified. However, with family-based studies, we are in the unique situation that we can estimate a_k from the data without biasing the test results. Because we will later compute the FBAT-GEE statistic that uses the offspring marker scores in the informative families, we can use only equation (1) for parameter estimation when the family is noninformative (both parents are homozygous) and is consequently not included in the computation of the FBAT statistic. Estimation of the model parameters in equation (1) on the sole basis of noninformative families is problematic for two reasons. First, since the number of noninformative families

may be small compared with the total number of families, the mean parameter estimates will not be very efficient when estimated solely on the basis of the data of the noninformative families. Furthermore, since the parents in noninformative families are homozygous, their offsprings' genotypes are also predominantly either 0 or 2, unless the allele frequency is ~0.5. The distribution of the marker scores used for the estimation of the genetic-effect sizes would therefore be highly skewed to either the left or the right. This extreme skewedness can make the estimation of the mean parameters unstable, and phenotypic outliers become highly influential.

To permit the use of both informative and noninformative families for the effect-size estimation without biasing the resulting test statistic, we replace the marker score x_{ij} in equation (1) by its expected value conditional on the parental genotypes $E(X_{ij}|p_{i_1},p_{i_2})$,

$$E(Y_{ijk}|p_{i_1},p_{i_2}) = \mu_k + a_k E(X_{ij}|p_{i_1},p_{i_2}) \ . \qquad (3)$$

It is important to note that equation (3) simplifies to (1) when the family is noninformative, because then the observed marker score x_{ijk} and the expected marker $E(X_{ij}|p_{i_1},p_{i_2})$ are identical, that is, $x_{ij} = E(X_{ij}|p_{i_1},p_{i_2})$.

Since the test statistic is based on the use of offspring genotypes conditional on parental genotypes, the use of equation (1) to estimate a_k does not bias subsequent testing, even in the informative families.

It should be noted that, for informative families, replacing the offspring genotypes by the conditional mean of the parental genotypes implies that the variance assumption changes, that is, $\mathrm{Var}(\mathbf{Y}_i|p_{i_1},p_{i_2}) = \mathbf{V}_i + \mathrm{Var}(a\mathbf{X}_i|p_{i_1},p_{i_2})$. Since it is common practice to assume that the genetic variance $\mathrm{Var}(a\mathbf{X}_i|p_{i_2},p_{i_2})$ is smaller than $\mathbf{V}_i$ (Falconer and Mackay 1997), the genetic variance is usually ignored (Martínez and Curnow 1992; Lange and Whittaker 2001). We disregard the genetic variance in our model specifications as well and assume that the phenotypic variance is given by equation (2), $\mathrm{Var}(\mathbf{Y}_i|p_{i_1},p_{i_2}) = \mathbf{V}_i$. We will estimate all parameters in equations (3) and (2) by the GEE approach (Liang and Zeger 1995). This method for estimating a_k and μ_k is robust against misspecification of the variance assumption. The robustness of the GEE approach is also important, because the specification of the variance matrix $\mathbf{V}_i$, which describes the phenotypic correlation within families and within individual offspring, is easily susceptible to errors. Using the robust effect-size estimates, we compute the conditional power of FBAT-GEE for the selected phenotypes, as outlined in Lange and Laird (2002*a*).

Thus, we propose the following algorithm to find the subset of phenotypes for which FBAT-GEE has optimal power:

- Select a subset of phenotypes.
- Posit a multivariate model that describes the phenotypes as a function of the genotypes.
- Replace the observed offspring genotypes by the conditional marker mean given the parental genotypes, that is, $E(X_{ij}|p_{i_1},p_{i_2})$.
- Estimate the genetic effect sizes, using the GEE approach for equations (2) and (3).
- Compute the conditional power (i.e., the power of FBAT-GEE given the observed data) for the multivariate FBAT-GEE test that uses the selected set of phenotypes.
- Repeat these steps until the group of phenotypes that has the highest power is found and then test this subgroup for association, by use of the multivariate FBAT-GEE.

It is important to note that a significant association between the phenotypic data and the marker locus is determined solely by the *P* value of the FBAT-GEE statistic, which is computed in the last step of the algorithm, using the actually observed offspring genotypes. Hence, our procedure is robust against population admixture and stratification. Nevertheless, its efficiency will be influenced by these effects.

Since the number of possible subsets of phenotypes grows exponentially with the number of observed phenotypes, it will often not be feasible to compute the power for all possible subsets. For applications to real data sets, one might elect an upper limit for the number of phenotypes for which the power of FBAT-GEE is estimated. An alternative is to use forward or backward selection strategies in the way that they are typically used in model building for regression analysis. For example, a potential forward selection strategy could be described by the following algorithm: Initially select the single phenotype with the highest estimated power. Then, in each subsequent iteration, add the phenotype to those already selected, which increases the power of FBAT-GEE the most. This is repeated until the power can not be further increased.

When the *P* value of the multivariate FBAT-GEE test, computed at the end of the algorithm, is significant for the selected group of phenotypes, all selected phenotypes must be tested individually by univariate FBATs. Since the multivariate FBAT-GEE and the univariate FBATs are applied sequentially and conditional on FBAT-GEE being significant, an overall significance level α can be obtained, when the same significance level α is applied in each step (i.e., first for the computation of FBAT-GEE and then for the computation of all univariate FBATs within the selected group of phenotypes). Thus, the initial FBAT-GEE which tests for the presence of an effect within the group of selected phenotypes, does not require any adjustment for multiple testing. However, when we test the selected phenotypes individually, the univariate FBATs must be adjusted for mul-

Table 1

Data Analysis for IL13

Method of Phenotype Selection[a]	FBAT-GEE	*P*	Power	$\hat{h}^2$	FBAT	*P*	Power
New testing strategy (group size 3):	11.54	.009	.99				
Total eosinophil count				.015	7.40	.006	.57
Post bronchodilator FEV 1% predicted				.002	.92	.34	.14
Albuterol use when exercising				.001	.16	.69	.08
DeMeo et al. (2002) (atopy group):	12.2	.0069	.52				
Total eosinophil count				.015	7.40	.006	.57
Total serum IgE				.005	3.6	.06	.21
No. of positive skin tests				.011	.21	.64	.45

NOTE.—Note that the power estimates for the univariate FBATs are not adjusted for multiple comparisons. To adjust for multiple testing, they must be computed for a significance level of $\alpha = .05/3 = .013$.

[a] FEV = forced expiratory volume.

tiple comparisons. It is important to note that this adjustment is not as strict as when testing all available phenotypes by univariate FBATs, since the number of selected phenotypes is usually much smaller than the total number of recorded phenotypes.

It is worthwhile noting that this applies only to the FBAT-GEE testing of the composite null hypothesis of the selected phenotypes. If the researcher wants to test additional hypotheses, this will again entail a problem of multiple testing.

Data Analysis: Childhood Asthma Management Program

We applied our new approach to phenotype selection to a collection of parent/child trios in the Childhood Asthma Management Program (CAMP) Genetics Ancillary Study. In the CAMP study, asthmatic children were randomly assigned to receive one of three different asthma treatments (CAMP 1999). Blood samples for DNA were collected from 696 complete parent/child trios from 640 nuclear families in the CAMP Ancillary Genetics Study. Baseline phenotype values, before randomization to treatment groups, were used in this analysis. Genotyping was performed at a polymorphism located in the interleukin (IL) 13 gene. For our analysis, we selected 22 phenotypes that have been analyzed elsewhere (DeMeo et al. 2002); as discussed in the earlier article, the use of univariate testing and adjustment for multiple comparisons does not lead to an overall significant result. DeMeo et al. therefore grouped the phenotypes into symptom groups and tested all phenotypes within one group simultaneously by the multivariate FBAT-GEE. When this strategy was used, the atopy group (including total eosinophil count, total serum IgE, and number of positive skin tests) showed an overall significant result. Individual testing within the atopy group indicated a significant association between the SNP in the IL13 gene and total eosinophil count. Since the grouping of phenotypes in this analysis was based on the investigators' clinical intuition, we will explore here whether our data-driven testing strategy would have led to the same result.

Assuming an additive model and a significance level of 5%, we applied our proposed testing strategy to 22 phenotypes. Since there are 2^{22} (4,194,304) subgroups of phenotypes, the calculation of the conditional power for all 4,194,304 subgroups of phenotypes is computationally intensive. We therefore decided to use the above-described forward-selection approach to select the phenotypes. Forward selection will also be used in the simulation study. Table 1 shows the phenotypes that have been selected by our testing strategy, the FBAT-GEE result, and the estimated power of FBAT-GEE. Furthermore, for each trait, the estimated effect size, the univariate FBAT result, and the power of the univariate FBAT are given. These values are not adjusted for multiple comparisons. All these quantities are also shown for the atopy group selected by DeMeo et al. (2002).

The most important observation in table 1 is that the variable total eosinophil count, which was the only trait for which DeMeo et al. (2002) could establish overall significance, is also selected by our testing strategy. Since the FBAT-GEE result for the selected group is significant ($P = .009$) at the 5% level, one can test all phenotypes of the group individually at an adjusted significance level of $\alpha = 0.05/3 = 0.017$. Again, the only variable that reaches overall significance is total eosinophil count. Thus, both analyses—the one based on symptom group (DeMeo et al. 2002) and the testing strategy proposed here—yield the same result.

The quantitative transmission/disequilibrium test (QTDT) proposed by Abecasis et al. (2000) is the only family-based association test that uses the noninformative families for the computation of the test statistic. Lange et al. (2002) reported a *P* value of .010 for the QTDT for this marker locus and total eosinophil count. Under the assumption of a univariate strategy testing all 22 phenotypes and with adjustment for 22 comparisons, this *P* value does not reach overall significance.

Furthermore, it is important to note that, by looking

Table 2

Simulation Study: Estimated Power Levels for Significance Level $\alpha = .01$

No. of Traits and Allele Frequency	Estimated Power (Maximal Power) when					
	$h = .025$		$h = .050$		$h = .100$	
	MI	MII	MI	MII	MI	MII
22:						
.05	.04	.10 (.21)	.11	.35 (.44)	.33	.70 (.78)
.10	.03	.10 (.23)	.16	.38 (.50)	.50	.76 (.84)
.20	.04	.11 (.24)	.16	.39 (.53)	.57	.80 (.87)
11:						
.05	.06	.10 (.21)	.12	.38 (.44)	.40	.71 (.78)
.10	.06	.11 (.23)	.18	.40 (.50)	.55	.78 (.84)
.20	.05	.11 (.24)	.21	.42 (.53)	.62	.81 (.87)
5:						
.05	.10	.14 (.21)	.23	.40 (.44)	.53	.72 (.78)
.10	.11	.15 (.23)	.28	.43 (.50)	.68	.79 (.84)
.20	.10	.16 (.24)	.32	.48 (.53)	.75	.84 (.87)

Note.—h = heritability; MI = univariate testing that is adjusted for multiple testing; MII = the new testing strategy. Maximal power levels are those that could be achieved when only the marker locus that is associated with the phenotype is tested. These values were obtained by using the approach described elsewhere (Lange et al. 2002).

at the power for three univariate FBATs, phenotypes can be tested at a higher power level by FBAT-GEE than by univariate FBATs. By combining both techniques (FBAT-GEE and conditional power calculations) and thereby using all the available data, it is possible to group phenotypes with low-powered univariate tests so that the overall FBAT-GEE statistic has sufficient power to detect a potential association.

Simulation Study: Power Comparison between Testing the Most Powerful Phenotypes by FBAT-GEE and Applying Univariate FBATs with Bonferroni Correction to All Phenotypes

In this section, we assess the power of the proposed testing strategy by using simulation experiments designed around the asthma study described in the previous section. As a basis for comparison, we compare our strategy with a Bonferroni approach that uses both a separate statistic and an adjusted α level for each phenotype. Although it is appropriate to use the univariate FBAT for quantitative traits, we use the QTDT because it also uses information from noninformative families. We will assess the effects exerted on the power by a variety of influences: allele frequency, heritability, the number of phenotypes analyzed, population admixture, and population stratification.

We assume that a sample consisting of asthmatics and their parents is given, and we observe the genotypes of all family members. As in the asthma study analyzed in the previous section, we assume that, for each offspring, 22 quantitative phenotypes are measured, and we want to test each phenotype for a potential association with the marker locus.

The trios with a biallelic marker locus are generated by drawing the parental genotypes p_1 and p_2 from the binomial distribution Bi(2,p), where p is the allele frequency of the disease gene in the population. We then use simulated Mendelian transmissions to generate the individual genotype x_i. The phenotypic vector $\mathbf{Y}_i$ for each offspring is a random sample from a multivariate normal distribution, that is, $\mathbf{Y}_i \sim N([a_1 x_i, \ldots, a_{22} x_i], \mathbf{V})$, where a_k is the additive effect for the kth phenotype and $\mathbf{V} = (\sigma^2_{k,l})$ is the (22 × 22) variance matrix. We measure the strength of the additive effect relative to the phenotypic variance by the heritability h^2_k (Falconer and Mackay 1997), which is the proportion of phenotypic variation explained by the genetic variation—that is, $h^2_k = \mathrm{Var}(a_k X_i)/\mathrm{Var}(Y_{ik})$. This expression for h^2 can be solved for a_k (Lange and Laird 2002*a*). We assume that the environmental variance for each phenotype is 1 and the environmental correlation matrix $\mathbf{V}$ is given by the environmental correlation matrix of the asthma study discussed in the previous section. The environmental correlation values of this matrix are shown in figure 1.

In each replicate of the simulation study, we generate 300 trios in which each offspring has 22 phenotypes. A value of 1% is selected as the overall significance level. For the 22 phenotypes, we assume that only one phenotype ($a_{j_0} \neq 0$) is associated with the marker locus and that the other phenotypes are not associated with the locus ($a_j = 0$). To compare the performance of the proposed testing strategy with the standard methodology, we test each phenotype by QTDT and adjust the P value for multiple comparisons by the Bonferroni correction (i.e., 1%/22). Elsewhere (Lange et al. 2003), we also used the Hochberg correction and permutation tests to handle multiple testing. Because these methods did not show a useful improvement over the Bonferroni correction and because they are computationally much slower (permutation tests), we omit these methods here. Instead, we apply our new testing strategy for multiple phenotypes, and we use the computationally fast forward-selection approach.

The selected phenotypes are tested for association with FBAT-GEE at a 1% significance level. If FBAT-GEE shows a significant result, all phenotypes of the group are tested individually by univariate FBATs, which are adjusted for multiple testing within the group of selected phenotypes. It is important to note that, because only one FBAT-GEE is initially computed, no adjustment for multiple comparisons is needed for the P value of FBAT-GEE.

For each alternative hypothesis $H_A: a_{j_0} \neq 0, j_0 = 1, \ldots, 22$, we repeated the simulation study 100,000 times. The power for each testing strategy was estimated by the proportion of the number of times the single phenotype associated with the locus is declared sig-

nificant. The averages of the power estimates for the 22 phenotypes/alternative hypotheses $H_A: a_{j_0} \neq 0$, $j_0 = 1, \ldots, 22$ are shown in table 2. By use of the approach to power calculations described elsewhere (Lange et al. (2002), table 2 also shows the power of the FBATs under the assumption that only the phenotype associated with the marker locus is tested. Since this is the optimal-case scenario, these power calculations are an upper limit for our testing strategy. Table 2 also shows the estimated power levels when, instead of 22 phenotypes, 11 and 5 phenotypes are analyzed. To estimate the significance levels for both approaches, we repeated the simulation study under the assumption that no trait is associated with the marker locus. The estimated significance levels for analysis of 22 phenotypes are shown in table 3.

Tables 2 and 3 show clearly that our new testing strategy can be substantially more powerful than univariate testing with Bonferroni correction, especially for low allele frequencies, small heritabilities, and cases in which many phenotypes must be tested. The power of our testing strategy increases slightly when the allele frequency is increased. Since, for higher allele frequencies, the number of noninformative families becomes smaller, this observation clearly illustrates the importance of also including the informative families in the effect-size estimation (equation [1]). Table 2 shows that reducing the number of phenotypes increases the power of our testing strategy only slightly; however, when the number of phenotypes is reduced, the differences between univariate testing adjusted by Bonferroni correction and our testing strategy become smaller. Nevertheless, it is worth noting that, even when only 5 phenotypes are analyzed, the advantages of our testing strategy are still of practical relevance. It is even more important to note that the power of our strategy decreases only moderately when the number of phenotypes from is increased 5 to 22. Comparing the power achieved by our testing strategy with the maximal possible power obtained by the approach by Lange et al. (2002), we observe that our testing strategy is 5%–14% less powerful than the optimal-case scenario. All these observations suggest that our testing strategy is especially powerful when many phenotypes are observed and we do not know the phenotype with the strongest genetic component for that candidate locus prior to the analysis.

The estimated significance levels indicate that our testing strategy is still too conservative, which is most likely caused by the Bonferroni corrections for the univariate

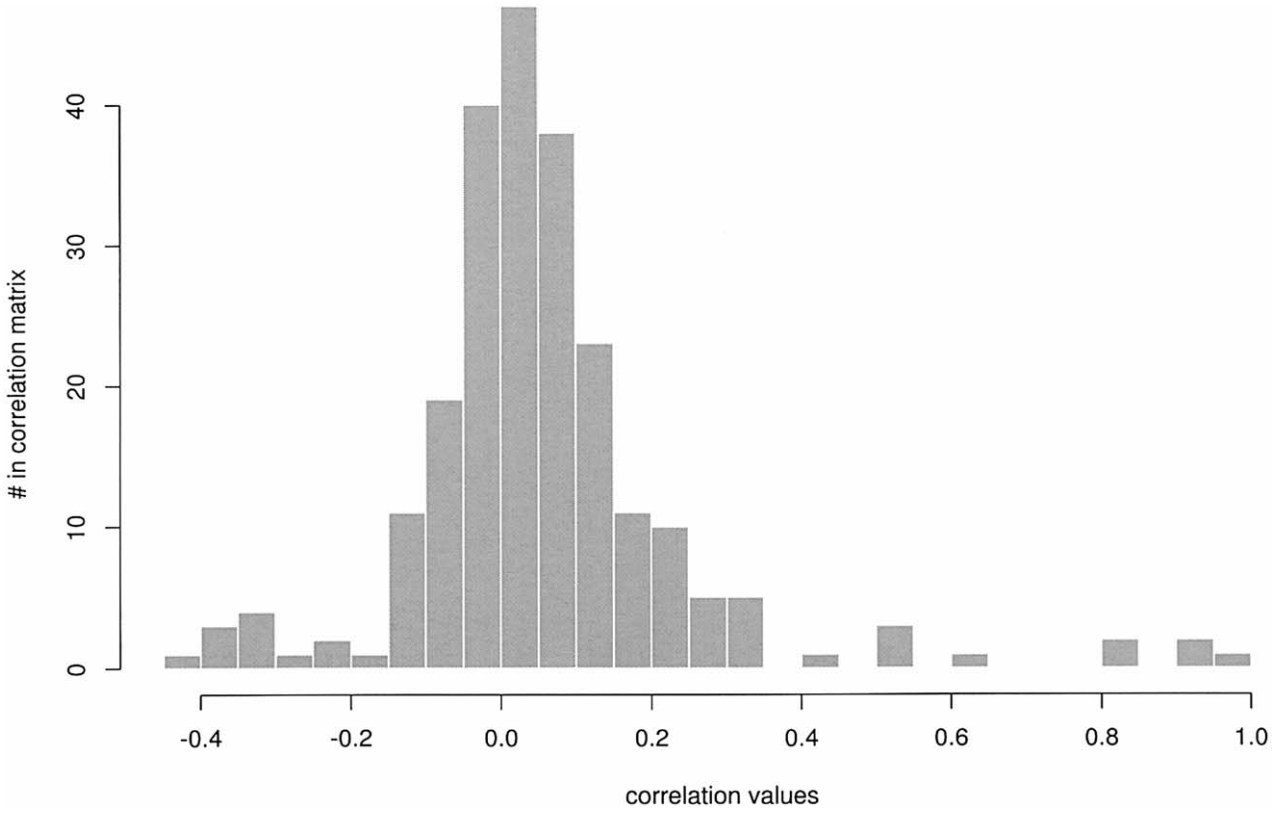

Figure 1 Histogram of environmental correlations

Table 3

Simulation Study for 22 Phenotypes: Estimated Significance Levels for the New Testing Strategy for Nominal Significance Level $\alpha = .01$

Allele Frequency	Estimated Significance Levels for	
	MI	MII
.05	.00019	.0030
.10	.00032	.0030
.20	.00020	.0021

NOTE.—MI = univariate testing that is adjusted for multiple testing; MII = the new testing strategy.

FBATs when testing within the group of selected phenotypes (table 3). On average, our testing strategy selects three phenotypes to be tested with FBAT-GEE.

Population Admixture and Stratification

We repeated the simulation study for scenarios with population admixture and stratification. As described elsewhere (Abecasis et al. 2000), we generated the population admixture by mixing two distinct populations of equal size. For each subpopulation, we assumed different allele frequencies and different phenotypic means (i.e., admixture and stratification). The allele frequencies in each subpopulation were constructed by adding and subtracting a certain percentage from the original allele frequencies in the first simulation study, that is, $p + x\%$ and $p - x\%$. The strength of the stratification is measured by the proportion of phenotypic variance due to the different phenotypic means in both subpopulations (Abecasis et al. 2000). The means in the two subpopulations were selected so that the subpopulation with the smaller allele frequency has the higher phenotypic mean. For a population-based analysis, this stratification would suggest a trend in the direction opposite to that of the true underlying additive effect, under the assumption that $a_i > 0$.

Assuming that the sample size is 300 and that 22 phenotypes are observed, we repeated the simulation study for a variety of values for population admixture and stratification. For univariate testing adjusted for multiple comparisons and for our new testing strategy, table 4 lists the estimated significance level for a nominal α level of .01. Table 4 shows that, in the presence of population admixture and stratification, the nominal α level is maintained for our new testing strategy, which is still too conservative. This observation confirms our theoretical conclusion that our testing strategy is robust against such effects.

The estimated power levels (table 5) suggest that our new testing strategy outperforms univariate testing, even in the presence of population admixture and stratification. For most scenarios, the advantages are substantial (gain of power of as much as 30%). However, for some scenarios, the estimated power levels of the two approaches become virtually identical. Table 5 suggests that, in general, one will not lose power by using our new testing strategy, even in the presence of population admixture and stratification.

The observation that our new testing strategy performs very well can be explained by the way the phenotypes are selected. The optimal combination of phenotypes is defined not by the absolute estimate for its power but through its ranking compared with the other groups of phenotypes. It seems that, because the other groups of phenotypes are also affected by admixture and stratification, the ranking based on the estimated power

Table 4

Simulation Study: Estimated Significance Levels for the New Testing Strategy for Nominal Significance Level $\alpha = .01$

Allele Frequency	Estimated Significance when							
	$s = .025$		$s = .50$		$s = .100$		$s = .200$	
	MI	MII	MI	MII	MI	MII	MI	MII
.05 ± 25%/2	.00034	.00049	.00026	.00049	.00024	.00049	.00028	.00049
.10 ± 25%/2	.00024	.00049	.00029	.00049	.00036	.00049	.00029	.00050
.20 ± 25%/2	.00022	.00049	.00027	.00049	.00027	.00049	.00024	.00048
.05 ± 50%/2	.00027	.00050	.00021	.00049	.00027	.00051	.00030	.00049
.10 ± 50%/2	.00031	.00049	.00029	.00049	.00017	.00049	.00027	.00099
.20 ± 50%/2	.00031	.00049	.00022	.00049	.00023	.00099	.00026	.00149
.05 ± 100%/2	.00026	.00099	.00017	.00080	.00020	.00049	.00023	.00099
.10 ± 100%/2	.00029	.00099	.00030	.00099	.00031	.00149	.00022	.00199
.20 ± 100%/2	.00022	.00109	.00020	.00099	.00025	.00241	.00014	.00199

NOTE.—s = strength of population stratification; MI = univariate testing that is adjusted for multiple testing; MII = the new testing strategy.

Table 5

Simulation Study: Estimated Power When the Significance Level Is Set at $\alpha = .01$

	Estimated Power Levels when																							
	$s = .025$						$s = .050$						$s = .100$						$s = .200$					
	$h = .025$		$h = .050$		$h = .100$		$h = .025$		$h = .050$		$h = .100$		$h = .025$		$h = .050$		$h = .100$		$h = .025$		$h = .050$		$h = .100$	
Allele Frequency	MI	MII	MI	MII	MI	MII	MI	MII	MI	MII	MI	MII	MI	MII	MI	MII	MI	MII	MI	MII	MI	MII	MI	MII
.05 ± 25%/2	.02	.07	.08	.25	.28	.55	.02	.06	.07	.19	.29	.53	.01	.03	.08	.17	.23	.45	.02	.05	.09	.21	.43	.62
.10 ± 25%/2	.02	.07	.13	.25	.41	.65	.02	.05	.10	.19	.41	.60	.01	.03	.09	.17	.39	.55	.03	.06	.11	.24	.44	.70
.20 ± 25%/2	.03	.08	.15	.26	.45	.70	.04	.04	.13	.18	.46	.63	.02	.03	.10	.16	.46	.56	.03	.08	.11	.28	.44	.77
.05 ± 50%/2	.01	.05	.06	.22	.24	.52	.02	.05	.07	.19	.23	.51	.01	.03	.07	.14	.20	.41	.02	.05	.08	.20	.39	.58
.10 ± 50%/2	.02	.05	.11	.22	.38	.59	.02	.04	.10	.19	.38	.54	.01	.03	.08	.13	.33	.48	.02	.07	.10	.24	.39	.67
.20 ± 50%/2	.03	.07	.14	.21	.42	.65	.02	.03	.12	.18	.41	.56	.01	.02	.09	.12	.38	.49	.02	.08	.10	.28	.42	.74
.05 ± 100%/2	.01	.02	.04	.14	.15	.42	.01	.04	.04	.12	.16	.38	.01	.02	.04	.09	.15	.33	.01	.03	.07	.18	.31	.46
.10 ± 100%/2	.02	.02	.07	.15	.29	.47	.01	.02	.07	.11	.29	.41	.01	.02	.06	.08	.28	.31	.02	.08	.07	.23	.29	.57
.20 ± 100%/2	.02	.03	.10	.15	.30	.49	.02	.02	.08	.11	.32	.41	.01	.01	.06	.07	.29	.30	.02	.09	.08	.25	.27	.66

Note.—s = strength of population stratification; h = heritability; MI = univariate testing that is adjusted for multiple testing; MII = the new testing strategy.

Table 6

Simulation Study: Estimated Power Levels for Significance Level $\alpha = .01$ when the Sample Is Obtained through an Ascertainment Condition

Ascertainment Condition, Trait Used, and Allele Frequency	Estimated Power Levels when $h = .025$		$h = .050$		$h = .100$	
	MI	*MII*	*MI*	MII	*MI*	*MII*
Lower 5% and upper 5%:						
Associated						
.05	.77	.60	.97	.82	1.00	.85
.10	.81	.63	.96	.93	1.00	.94
.20	.81	.63	.98	.98	1.00	.98
Not associated						
.05	.06	.09	.11	.28	.37	.61
.10	.07	.10	.16	.31	.49	.68
.20	.07	.11	.18	.32	.50	.72
Upper 10%:						
Associated						
.05	.37	.26	.64	.49	.95	.51
.10	.39	.34	.80	.52	1.00	.53
.20	.45	.45	.90	.56	1.00	.96
Not associated						
.05	.04	.08	.12	.27	.38	.67
.10	.05	.08	.17	.28	.50	.70
.20	.05	.09	.18	.29	.51	.71

Note.—h = heritability; MI = univariate testing that is adjusted for multiple testing; MII = the new testing strategy.

of the phenotype groups is relatively well preserved in the presence of admixture and stratification. However, the actual estimates of the power may be biased.

Ascertained Samples

Finally, we repeated the simulation study for samples that are obtained through an ascertainment condition. Here, we assumed that population admixture and stratification are absent and that 22 phenotypes are observed for each offspring. First, the simulation study was repeated under the assumption that the researcher correctly anticipates the phenotype that is associated with the marker locus and ascertains the data set by sampling in equal parts from the upper 5% tail and the lower 5% tail of the distribution of this phenotype. Both testing strategies are then applied to the 22 phenotypes of the ascertained sample. The power levels, estimated on the basis of 100,000 replicates, are shown in table 6.

The same simulation studies were then repeated under the assumption that the researcher selects a phenotype that is not associated with the marker locus and uses this phenotype to ascertain the sample as before. In the simulation study, we mimicked this "unlucky" choice by randomly selecting a phenotype from the 21 phenotypes that are not associated with the marker locus and using this phenotype in the ascertainment condition. The estimated power levels, based on 100,000 replicates, for this scenario are also given in table 6.

The results in table 6 clearly illustrate that, if the researcher ascertains the samples through an ascertainment condition for the trait that is associated with the marker locus, univariate testing adjusted for multiple testing can be substantially more powerful than our testing strategy. However, when the researcher is less fortunate in selecting the phenotype and elects a phenotype for the ascertainment condition that is not in association with the locus, our new testing strategy again dramatically outperforms univariate testing. For this scenario, the estimated power levels of the new testing strategy are only slightly lower than those for total population samples (table 2). It is also worth noting that in a real-life situation one would not assign equal α levels to each phenotype in the univariate testing strategy (i.e., $\alpha/22$) but would emphasize the selected trait (e.g., $0.9 \times \alpha$) for the selected trait and $(0.1 \times \alpha/21)$ for the remaining 21 phenotypes. For all scenarios shown in table 6, such weighted α levels will make the differences between both testing strategies even greater.

To assess the effects of different ascertainment conditions on the power of both testing strategies, the simulation studies for both scenarios were repeated under the assumption that only the upper 10% tail of the phenotypic distribution of the selected trait is ascertained (table 6). Although the estimated power levels are, in general, slightly lower for this ascertainment condition ($Y > 10\%$), they do not differ substantially from the results for the first ascertainment condition ($Y > 5\%$ or $Y < 5\%$).

Therefore, in practice, before the data are collected, researchers must decide how confident they are of being able to identify the correct phenotype for the ascertainment condition. If the investigator strongly believes that the correct phenotype can be identified, ascertaining the data through an ascertainment condition on the identified phenotype and using univariate tests will be more powerful than our testing strategy. However, when researchers do not have a good intuition about the phenotypes with the strongest genetic component, they may find it less risky to collect a total population sample with a set of relevant phenotypes. This strategy has the additional advantages that the financial resources used in the screening process to ascertain a sample can be redirected and used to obtain an even bigger total population sample, which will further increase the power of our new testing strategy.

Discussion

In the present article, we propose a new sequential testing strategy for family-based association tests when multiple phenotypes are recorded. The approach allows us to screen all possible null hypotheses without biasing the overall significance level and to select a group of phe-

notypes that can be tested with maximal power. Although the approach uses all the available data, the noninformative families as well as the informative ones, it is robust against population admixture and stratification. The approach is relatively simple, takes advantage of all available data, even the noninformative families, requires no adjustments for multiple testing, dramatically outperforms univariate testing with Bonferroni correction, and is fairly robust against population admixture and stratification. Researchers can use our approach whenever many phenotypes are recorded and it is not obvious which phenotypes are most relevant for the marker locus.

In general, the success of our testing strategy will depend upon how well equations (2) and (3) describe the phenotypic data. Since the parameters in equation (2) and (3) can be estimated as many times as wanted without biasing the nominal significance level of the FBAT-GEE statistic, it is recommended to do statistical model building for equations (2) and (3) before finally estimating the power of FBAT-GEE. For example, when other variables are known to have an influence on the modeled phenotype/trait, these variables should be included in the model equation for the mean (1) as covariates, to avoid confounding of the estimates for the genetic effect sizes (Lange et al., in press). Also, different modes of inheritance (e.g., dominant or recessive) should be explored in equations (2) and (3) to prevent biased estimates for the power of FBAT-GEE. Furthermore, if the data are not normally distributed, one can use multivariate models designed for nonnormal data (Prentice and Zhao 1991) to model the phenotypes as a function of the conditional marker score.

The main approach of the present article, that is, conditional power calculations using genetic effects estimated by the biometrical model for the phenotypes where the offspring genotypes are replaced by their expected values conditional on the parental genotypes, can be used for a variety of other important applications. Since these power calculations can be applied to a set of marker loci (e.g., scans of candidate genes), they can be used to detect loci that have high power when tested for association with FBATs. Again, the data used to find these loci can later be used for the final association tests without biasing the nominal significance level (authors' work in progress). Furthermore, the proposed methodology can be used to include predictor variables in FBATs so that the power of the test statistic is maximized (authors' work in progress).

A nonparametric version of the proposed methodology is discussed elsewhere (Lange et al. in press). The approach has been implemented in our software package called "PBAT" and is available on the PBAT Web page.

Acknowledgments

We would like to thank two referees for their helpful comments on an earlier draft of the present article. We thank all the CAMP families for their enthusiastic participation in this study. We acknowledge the CAMP investigators and research teams for collecting the CAMP data. CAMP is supported by National Heart, Lung, and Blood Institute (NHLBI) contracts N01-NR-16044, 16045, 16046, 16047, 16048, 16049, 16050, 16051, and 16052. The CAMP Genetics Ancillary Study is supported by NHLBI grant P01 HL67664, and additional support for this research came from NHLBI grants R01 HL66386, T32HL67427, and HL66795. C.L. and N.M.L. were supported by NHLBI grants P01 HL67664, R01 NL66386, N01 HR16049, T32 HL07427, and N01 HLC6795. Comments from two referees were very helpful in preparing this version of the article.

Electronic-Database Information

The URL for data presented herein is as follows:

PBAT Web Page, http://www.biosun1.harvard.edu/~clange/pbat.htm (for PBAT software)

References

Abecasis GR, Cardon LR, Cookson WOC (2000) A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66:279–292

DeMeo DL, Lange C, Silverman EK, Senter JM, Drazen JM, Barth MJ, Laird NM, Weiss ST (2002) Univariate and multivariate family based analysis of the arg130gln polymorphism of the IL13 gene in the childhood asthma management program. Genet Epidemiol 23:335–348

Falconer DS, Mackay TFC (1997) Introduction to quantitative genetics. Longman, New York

Childhood Asthma Management Program Research Group (1999) The childhood asthma management program (CAMP): design, rationale, and methods. Control Clin Trials 20:91–120

Laird NM, Horvath S, Xu X (2000) Implementing a unified approach to family based tests of association. Genet Epidemiol Suppl 19:S36–S42

Lange C, DeMeo D, Laird NM (2002) Power and design considerations for a general class of family-based association tests: quantitative traits. Am J Hum Genet 71:1330–1341

Lange C, Laird NM (2002*a*) Analytical sample size and power calculations for a general class of family-based association tests: dichotomous traits. Am J Hum Genet 71:575–584

Lange C, Laird NM (2002*b*) On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power and optimality considerations. Genet Epidemiol 23:165–180

Lange C, Lyon H, DeMeo D, Raby B, Silverman EK, Weiss ST. A new powerful nonparametric two-stage testing strategy for family-based association tests for testing multiple phenotypes using all available data. Hum Hered, in press

Lange C, Silverman EK, Xu X, Weiss ST, Laird NM (2003) A

multivariate transmission disequilibrium test. Biostatistics 71:195–206

Lange C, Whittaker JC (2001) A generalized estimating equation approach to mapping of quantitative trait loci (QTL). Genetics 159:1325–1337

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22

Martínez O, Curnow RN (1992) Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor Appl Genet 85:480–488

Nelder JA, McCullagh P (1989) Generalized linear models, 2nd ed. Chapman and Hall, London

Prentice R, Zhao L (1991) Estimating equations for parameters in means and covariance of multivariate discrete and continuous response. Biometrics 47:825–839

Rabinowitz D, Laird NM (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered 50:211–223

Risch N, Merikangas (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516

Thomson G (1995) Analysis of complex human genetic traits: an ordered-notation method and new tests for mode of inheritance. Am J Hum Genet 57:474–486

Zhao H (2000) Family-based association studies. Stat Methods Med Res 9:563–587